探索 Iris（鸢尾花）数据集

CS245　数据科学基础　陆朝俊
叶泽林 515030910468

1　问题描述

统计学主要研究事物的数量方面，目的是探索数据集的数量特征。而统计学中的描述统计学借助图表或概括性的数值将数据集展示为清晰可理解的形式。在之前的研究中，已经使用了图表对 Adults 数据集进行了探索和可视化展示，这次研究的主要目标是探索 Iris 数据集，并通过一些概括性的数值对其进行展示，详细目标如下：

1. 探索 Iris 数据集的基本属性（如数据集总体描述、数据维数、特征名称等）；
2. 探索各特征的最小值、最大值、均值、中位数、标准差；
3. 探索各特征之间，以及特征与目标之间的相关性（相关系数）

2　解决方案

3　结果展示

3.1　Iris 数据集基本属性

表 1: Iris 数据集基本属性

属性	值
实例的数据类型	numpy.ndarray
实例的数据维数	(150, 4)
特征名	萼片长度，萼片宽度，花瓣长度，花瓣宽度 (单位均为 cm)
实例的类别值	0,1,2
实例的类别维数	(150,)
类别名称	setosa(清风藤), versicolor(云芝), virginica(锦葵)

3.2　各特征的数值描述

3.3　各特征及特征与目标之间的相关性

表 2: Iris 数据集各特征的数值描述*

	萼片长度	萼片宽度	花瓣长度	花瓣宽度
最小值	4.3	2.0	1.0	0.1
最大值	7.9	4.4	6.9	2.5
均值	5.843	3.054	3.759	1.120
中位数	5.80	3.00	4.35	1.30
标准差	0.825	0.432	1.759	0.761
方差	0.681	0.187	3.092	0.579
极差	3.6	2.4	5.9	2.4
下四分位数	5.1	2.8	1.6	0.3
上四分位数	6.4	3.3	5.1	1.8

表 3: Iris 数据集各特征及特征与目标之间的相关性

	萼片长度	萼片宽度	花瓣长度	花瓣宽度
萼片长度	1.000	-0.109	0.872	0.818
萼片宽度	-0.109	1.000	-0.421	0.357
花瓣长度	0.872	-0.421	1.000	0.963
花瓣宽度	0.818	-0.357	0.963	1.000
目标	0.783	-0.420	0.949	0.956